Cent. Eur. J. Biol. • 8(5) • 2013 • 399-409
DOI: 10.2478/s11535-013-0155-8

VERSITA

Central European Journal of **Biology**

A practical introduction to microbial community sequencing

Mini-Review

Dominika Chmolowska*

Institute of Environmental Sciences,
Jagiellonian University,
30-387 Kraków, Poland

Received 18 September 2012; Accepted 24 January 2013

Abstract: The use of molecular methods is gaining popularity throughout the field of microbial community ecology studies thanks to their flexibility of application, which ranges from community structure to function and trait determination. Nonetheless, there are environmental microbiologists, who are new in the field and are just starting to get to grips with the genetic tool box. It is for them that this practitioner's mini-review was compiled. The methods available for microbial community structure analysis are discussed, after which, the reader is introduced to sequencing, as this tool is the most appropriate and has seen the greatest development in recent years. A focus on the practical aspects of the methodology is maintained throughout. The sample preparation procedure from extraction to sequencing is described. Different applications and considerations of sequencing are briefly explained, including clone library sequencing *vs*. amplicon library sequencing, shotgun-metagenomics *vs*. metatranscriptomics and the 'double RNA approach'.

Keywords: *Microbial community structure* • *Metagenomics*

1. Introduction

The great popularity of methods based on the analysis of genetic material which has unquestionably been observed over the last decades has come about as a result of their quickness, their flexibility of application and trends in research. Many scientists in fields other than molecular biology who have not yet added these methods to their lab tools are now considering doing so. It has now become a standard practice in research to use molecular tools in studying microbial communities. This article is intended for scientists whose experience of the genetic tool box has so far been limited. The main topic is sequencing, one of the molecular methods which can be used in the study of microbial communities. It is particularly useful in describing their biodiversity and has gained a great deal of attention in recent years.

1.1 Studying the structure of microbial communities

What are the methods available for studying the structure of microbial communities? They are based on either the cultivation of microbes *per se* or on biochemical analyses, in other words, the isolation and/or detection of their cell compounds, namely, lipids, proteins and nucleic acids.

1.1.1 Cultivation

Culture selections are widely disapproved of because only a limited fraction of the total cell number grows in an artificial environment; in the case of soil, about 5% of the culturability is commonly the upper limit [1]. Some researchers have argued, though, that the conditions established in order to achieve culturing need to be optimised [1]. With the use of diffusion chamber-based methods, average culturability has been increased fivefold, reaching approximately 35% as an average and 50% as a maximum of total seawater or soil bacteria, including many new taxonomic units. However, several important groups, such as *Actinomycetes* and *Acidobacteria*, were missing [2]. Cultivation is still more a tool for those who are interested in describing new species and wish to devote their energies to isolating single cell lines or high cell biomass is needed, eg. for biotransfomations. In Biolog plates, which were used to investigate functional diversity of communities [3], 31 to 95 carbon substrates are employed. Does this variety of substrates increase the actual culturability on Biolog

* *E-mail: dominika.chmolowska@uj.edu.pl*

plates? It remains a question since no study has been conducted into this to date.

1.1.2 Biochemical analyses

Phospholipid Fatty Acid analysis (PLFA) provides information about the relative biomass of different groups, namely, total bacteria, bacteria G(+) and G(-), total fungi, *Ascomycota*, archaea and protozoa. As a method, it is more or less limited to that level of resolution [4], with rare exceptions such as individual FA-markers for Arbuscular Mycorrhizal Fungi (AMF) [5] or some thermophylic *Alicyclobacteria* and methanothropic bacteria [6]. When PLFA are used for the between samples comparison of the biomass of different microbial groups, the analyses need to be treated with caution, since PLFA cell biomass is biased by the nutrient status of the microbial cells [4]. In addition, environmental conditions affect the contents of particular types of PLFA, which can be exploited in order to infer the physiological status of a population or community, such as, for example, when bacteria move from the logarithmic to the stationary phase and starvation monoenoic acids are converted to cyclopropane PLFA [4]. Lipid markers can also be used to identify a pure culture sample by comparing its PLFA profile against a database. Commercial packages such as the Sherlock® Microbial Identification System (MIS) are available for this purpose.

Analysis of the protein content, which is to say, the proteome, is less commonly used for studying community structure. However, it has recently been applied in several studies as a means of functional community profiling. The number of proteins present in microbial cells is considerable, with up to 18 000 different proteins being identified from a simple community composed of only six bacterial species [7]. This makes metaproteomics too difficult to translate to complex community composition and when such attempts were made, a lower resolution was applied [8]. To aid the identification of protein spectra, metaproteomic studies are often combined with shotgun metagenomics [9]. Protein markers and antibodies have also been used for species detection with reasonable success. Proteomic analyses are more expensive than genomic analyses and the information obtained is more difficult and laborious to analyse. On the other hand, it is a powerful tool in the study of community functioning. For example, changes in the population/community proteome are monitored before and after exposure to a stressor in order to establish the metabolic state.

Raman spectroscopy, which can be used to generate biochemical fingerprints which include nucleic acids, carbohydrates, lipids and proteins, has been employed for studying microorganisms in recent years. Confocal Raman microscopy has been shown to allow the rapid discrimination of bacteria species from colonies or even from single cells [10]. However, the use of this method in the study of bulk environmental samples can be limited. Like PLFA, it has been used in commercially available microbial identification systems [11].

1.1.3 Nucleic acid-based methods

The spectrum of analytical methods based on nucleic acids is wide. They rely on different procedures and result in different endpoints. They can be roughly grouped into: 1) quantification of cells, 2) microarrays, 3) fingertyping, 4) sequencing.

Whole cell methods such as staining with DNA-binding dyes (DAPI, SYBR®-Gold or SYBR®-Green) are useful for counting total cell numbers. In addition, Fluorescent *In Situ* Hybridization (FISH) and Catalyzed Reporter Deposition–FISH (CARD-FISH) allow the detection and quantification of the target population in question, for example, *Acidobacteria, Alfaproteobacteria,* and so forth; they also permit the researcher to construct a probe with a taxonomic resolution as low as is required. When combined with the incorporation of radio-labelled substrates, which is known as Microautoradiography-FISH (MAR-FISH), they can supply additional information as to which microbial group is metabolically active. However, this method is not free from biases, which result from the fact that there is no artificial substrate which provides uniform uptake by cells [12,13]. Microautoradiography evolved into another method, nano Second Ion Mass Spectrometry (nanoSIMS), which widens the isotope range and substantially improves the quantifiability [14]. As with whole cell methods, however, these, too, might prove to be laborious and time consuming if there are numerous samples to be analysed and the community is complex.

Quantitative, or real-time, PCR (qPCR) is somewhat similar to FISH. In as much as FISH supplies information on the number of target cells, qPCR provides data on the target gene copies in a sample. This can be employed for the quantitative comparison of phylogenetic groups or biochemical activities, for example, in order to posit the bacteria to archaea ratio in a sample [15]. Like FISH, it would be inapplicable to complex structure studies, especially given that there are more convenient methods.

Phylochips is a patented method based on microarray phylogenetic analysis. As with FISH, this is a hybridization-based method in which specific probes attached to a matrix are exposed to a template extracted from the sample. The hybridization between the template and the probe is quantified by fluorescence

and can provide information regarding the presence of a target group. Phylochips can be designed for a sample type, such as soil microbial community phylochips, for example, or for a target group, such as, for instance, proteobacteria phylochips. They may consist of several hundreds of thousands of probes [16]. This method is not labour-intensive and permits the community structure, together with information on its composition, to be obtained quickly. Similarly, the Geochips method offers a functional profile of the community transcriptome on the basis of mRNA probes [17]. The major disadvantage of microarray technology is its rather high cost. Furthermore, it might not show the full community structure; it has been reported that, in a complex community, the unique species were not covered [18,19]. It is worth mentioning about reverse line blot hybridization [20], an analogue method to phylogenetic microarrays, which is, though, less commercialised and automated. For that reason is would be rather applied, similarly to FISH, for detection of chosen phyla than to overall community composition determination.

Molecular fingertyping shows microbial community profiles, in most cases without an insight into their taxonomic composition. The analysis used to be carried out by means of the graphical examination of DNA fragment distribution on a gel. Nowadays, they are more commonly sorted on to a Sanger sequencer. Gradient gels such as Denaturing (DGGE) or Temperature Gradient Gel Electrophoresis (TGGE), which were popular at the turn of the 20th and 21st centuries, have recently been overtaken by genotyping and methods such as Terminal Fragment Length Polymorphism (T-RLFP), Amplified Ribosomal DNA Restriction Analysis (ARDRA) and Automated Ribosomal Intergenic Spacer Analysis (ARISA). This method can serve in the estimation of the richness, diversity and similarity of environmentally-derived samples. The profiles are constructed from amplified taxonomic marker genes, such as, for instance, 16S rDNA for bacteria and archaea. In the case of T-RLFP and ARDRA, PCR is followed by a restriction enzyme treatment, which cuts the DNA in enzyme-specific regions and create a collection of DNA pieces. These are sorted, on the basis of size, in a Sanger capillary sequencer or on a gel. The resulting graph is a mass spectrum, a fingerprint of the community. These methods are much cheaper and faster than sequencing, and they allow the comparison of general microbial community structures or the observation of changes in the community following treatment. T-RLFP kits, which are already available on the market, are equipped with fluorescently-labelled primers, nucleases and access to the commercial database of microbial T-RLFP profiles. For more details and a review on the genotyping of microbial communities see [21]; a review focused on T-RLFP analysis can be found in [22]. Nevertheless, molecular genotyping is an approximate method in which insight into the composition of the community is highly limited.

Several analytical strategies have evolved on the basis of sequencing, namely, clone-library sequencing, amplicon library sequencing, and shot-gun sequencing; then, at a further level, we have DNA sequencing or RNA sequencing, through reverse transcription into cDNA. A fragment of a gene, a gene, a genome or the whole (meta)genome obtained from a sample can be sequenced, depending on the question the research is setting out to answer and the sample type.

For the analysis of microbial communities, the most common procedure is the sequencing of the phylogenetically conserved part of a gene which is known as a ‘marker gene’. Typical phylogenetic marker genes are the small subunit 16S or 18S, ribosomal RNA, for prokaryotes and eukaryotes respectively, the interribosomal spacer: internal transcribed spacer (ITS) and intergenic spacer (IGS), typically used for fungi, and the beta subunit of the RNA Polymerase (*rpoB)* or cytochrome c oxidase, which is employed for eukaryotes. The sequence is compared to the database and the closest match is retrieved. If this match is assigned to a cultured specimen, further information regarding the metabolism and physiology can be obtained.

This profiling requires the amplification of the marker gene, which is achieved by means of a Polymerase Chain Reaction (PCR). In contrast, a cleaned, extracted, nucleic acid can also be directly subjected to sequencing without amplification. This PCR-independent method is called shot-gun sequencing on account of its rapidity. DNA and RNA can undergo sequencing from one sample. Analysis of the DNA provides information as to the total community in the environment, and the RNA, about its metabolically active fraction.

Molecular methods are considerably less laborious and faster than other methods; if one were to extract phospholipid fatty acids (PLFA) from soil, twenty samples would take from two days to two weeks depending on protocol. For extraction of soil DNA, only one to three days would be needed for the same number of samples. The high throughput is especially convenient when a high number of samples are to be compared. In good quality ecological studies, where due attention is paid to proper replicates and statistics [23], it often happens that the samples are counted in dozens and hundreds. Another consideration is the optimisation of the molecular pipeline solving a difficult template, which is often time-consuming and requires high levels of patience and financing.

The brief foregoing discussion and comparison of methods suggests that sequencing is a smart solution for the overall description of a microbial community's structure. A more detailed overview can be found in a book entitled *Biological diversity: frontiers in measurement and assessment* [24]. What follows is a guideline on how to prepare a sample from scratch.

2. Tools and procedures

2.1 Sample preservation

It is always best to process samples as soon as possible, but when tens or hundreds of samples are gathered over a short period of time, their preservation might be necessary. For the purposes of DNA extraction, soil can be frozen. Before freezing, water samples can be filtered through disc membranes or syringe filters with the appropriate mesh, for example, 0.2 µm pores if bacteria and archaea are to be targeted. This is very convenient for both transportation and increasing template concentration. However, one needs to be aware of filtration biasing [19]. If the samples are RNA target, they can be preserved in RNALater®, for instance, after a water sample has been filtered, syringe filters can be partially filled with this preparation, while sediments can be kept in plastic tubes filled with it.

2.2 Nucleic acid extraction

DNA or RNA can be extracted manually or with one of the several kits available on the market. Different sample preservation and extraction methods isolate different fractions of the total community [25-27]. A good quality extract is crucial to analytical success because, depending on the sample source, overpowering inhibitors, such as, for example, humic acids from soils or even the vestiges of extracting agents such as alcohols, phenol, chloroform, SDS or EDTA, can inhibit further applications.

The quality of the extract can be ascertained by several means. NanoDrop UV Spectrometer is widely used. A 1 µl drop elicits information from the spectra as regards the amount of DNA or RNA and contaminants. Absorbance on 260 nm provides information on nucleic acid quantity while 260 nm/280 nm and 260 nm/230 nm gives ratios of purity. 260 nm/280 nm =1.8 is considered as pure DNA, and 260 nm/280 nm =2.0 as pure RNA. If the ratio is different, it might indicate the presence of proteins, phenol and other contaminants. The ratio for 260 nm/230 nm can be a little higher than the one for 260 nm/280 nm and varies between 1.8–2.2. If the ratio differs greatly, it may point to the presence of phenol, guanidine, magnetic beads, carbohydrates and/or proteins. However, NanoDrop is not very accurate in measuring DNA concentrations of lower than 10 ng μl^{-1} and it is also often the case that the concentration measurements are exaggerated on account of the overlapping absorbance of contaminants.

DNA and RNA content can also be measured fluorometrically by means of Qubit or PicoGreen. These methods are more sensitive than NanoDrop. Qubit only provides information on template concentration, while PicoGreen, like NanoDrop, also provides a feedback on contaminants. In many protocols for amplification, restriction or sequencing, a template concentration is standardized for the concentration and using Qubit or PicoGreen rather than NanoDrop is suggested.

It is crucial to control nucleic acid fragment sizes during the preparation of any genetic material. DNA or RNA extracts are most popularly examined on a 0.9% agarose gel. GelRed or SybrSafe dyes may be used instead of toxic Ethidium Bromide. The thinner the gel, the sharper the bands will be; gels of around 5 mm should work very well. Loading 5 µl of the extract on to the electrophoresis should result in a visible band of around 23 kb in size. The more distinct the band, the better; at the same time, smearing below the band is caused by the migration of smaller nucleic acid pieces from fragmented DNA. In the case of a highly concentrated DNA extract, the product could be immersed in a well and still no band would be visible. Loading a smaller quantity can help in such instances. The concentration of a template can be roughly assumed by comparing the extract band to one of the standard bands with a similar intensity and calculating the total content of a fragment.

RNA extraction product can also be visualized on an agarose gel. This allows the verification of the approximate quantity of the three rRNA subunits, namely, the small subunit or SSU (16S in prokaryotes and 18S in eukaryotes), the large subunit or LSU (23S in prokaryotes or 28S in eukaryotes) and 5S (prokaryotes)/5.8S (eukaryotes); it also permits the mRNA to be checked. For this reason, RNA is usually denatured, since, rather than remaining single stranded it forms secondary structures. The simplest procedure is heating the sample for five minutes in formamide, then cooling on ice, followed by standard electrophoresis in TAE 1x, as in Masek *et al*. [28]. However, denaturation makes the bands fainter and there has to be a considerable amount of RNA in the extract in order for it to be seen on agarose. RNA can be also visualised without denaturation as bands of a smaller weight than DNA.

The best way of checking nucleic acid quality, particularly RNA, is by means of Bioanalyzer chips (Agilent), which are three-in-one. They integrate the

concentration, purity and integrity measurements. This method is much more financially demanding than the most commonly employed set, NanoDrop plus agarose, though.

DNA extracts can be kept at 4ºC for up to one week and frozen at -20ºC for months. RNA is very labile and it has to be frozen to a minimum of -20ºC as soon as possible. It is highly recommended to freeze it in liquid nitrogen directly after extraction is complete and then to store it at -70ºC.

Soil DNA extracts, when prepared in line with, for example, Blagodatskaya *et al*. [29] or Aoshima *et al*. [30], can be used for microbial biomass estimation. Apart from this, the extract is just a starting point for further protocols.

2.3 cDNA synthesis

Because RNA is very labile, it has to be used as a template to synthesize complementary DNA (cDNA) before any PCR-dependent or PCR-independent sequencing is carried out. To do this, the reaction mixture needs to be added, namely, reverse transcriptase, dNTP mix, and primers. Primers can be specific to the gene in question or they can be non-specific hexamers. For eukaryotes, polyA-primers targeting the polyadenylated tail of mRNA transcripts can also be used. It is wise to ascertain which set will give the better yield with a particular template. Single stranded (ss) or double stranded (ds) cDNA, followed by PCR, provides the material for sequencing. If there is a large quantity of RNA, then the shotgun sequencing of ds cDNA can be considered.

2.4 PCR

Polymerase Chain Reaction multiplies a marker gene for analytical methods in which its high content is necessary. PCR precedes clone library sequencing and amplicon library sequencing. For PCR, the components which are usually mixed are sterile milli-Q water, polymerase, the polymerase buffer, primers and template. Primers work as starters and determine which region of the nucleic acid will be amplified. This can be a part of a gene or the whole gene. A part of a gene is multiplied in the case of phylogenetic analysis. The most commonly used taxonomic marker genes are the genes coding the Small Subunit (SSU) of ribosomal RNA; in the case of bacteria and archaea, this is 16S. On rare occasions, the Large Subunit (LSU or 23S) can also be used for prokaryotes and it is said to provide better discrimination between closely related operational taxonomic units (OTU). LSU is a more common target in eukaryotes, where it is known as 28S. Fragments of 28S rDNA or inter-ribosmal transcribed region (ITS) are, on the other hand, a target for primers in the case of Eucaria such as fungi. Apart from this, genes encoding enzymes can be used as phylogenetic markers, for example, RNA polymerase (rpoB, RPB1), ammonium monooxygenase (AMO), nitrite reductase (nirK, nirS), cytochrome c (cyt c), as well as for the purpose of determining functional activities. When choosing a primer set, one should be aware of its coverage [e.g. 31-33] and possibly lowest preferential amplification [34]. The primer match can be checked at: www.rdp.cme.msu.edu or www.arb-silva.de. At more advanced level a researcher may consider designing his/her own primers, according to general rules, which in short are listed on many web-sites, in publications e.g. [35] or with details in books. It is suggested to use software for primer design and there are many of them reviewed, for example at http://www.molbiol-tools.ca/PCR.html.

It is important to use the optimal DNA concentration. Using too little might exclude the genomes of less abundant microorganisms and decrease the diversity measurement. Using too much can also inhibit PCR by way of a high DNA concentration or an increased amount of contaminants. The amplification of environmentally derived templates such as soil, sediments, or acid mine drainage is a challenge because of contaminants. The addition of BSA to a PCR of low purity extracts increases the yield. In the case of soil samples, the use of a concentration of around 0.5–1 μg BSA μl^{-1} of PCR mix is a common procedure. Other PCR enhancements are DMSO [36] and protein T4 [37]. Given the common difficulties with optimising the reaction conditions, it is highly recommended to include a positive control. The positive control is a template that amplifies with the chosen primers under the selected conditions. Should no PCR product be obtained from the extract, it allows the suspicion of the wrong PCR setting to be excluded and points to problems connected with that particular template in terms of concentration or contaminants. When the PCR setting is correct, in that the positive control yields a product, but there is no amplification from the extracts and it is suspected that this might have resulted from inhibition by contaminants, then what is known as a 'spiked PCR' can be prepared. In this procedure, a positive control is combined with the extract in one PCR tube. In the case of a high concentration of contaminants, the spiked positive control will not amplify. A negative control, which is to say, one that contains the same PCR mix as that used for the samples and no template, should be included in every PCR. It controls foreign DNA contamination.

The first step of PCR is to heat the sample in order to melt its double stranded structure. In theory, the temperature depends on the structure of the particular piece of DNA; in practice, 94-95°C usually proves to be

effective. This is followed by annealing, during which, the primers bind to the single strands; the temperature for this step is between 50-63°C and is chosen on the basis of the primers' melting temperature and verified empirically. The next step is elongation at 72°C, where the polymerase synthesizes the complementary strand. For longer PCR products, sufficient time elongation is crucial for completing of the DNA synthesis. The last three steps are repeated, for instance, twenty to thirty times, in order to provide a good amount of the product. Finally, the PCR mixtures are cooled to 8 or 4ºC to stop the reaction and preserve the product.

The relation of the number of cycles to the quantity of the amplicon produced should be adjusted to the lowest number of cycles possible. This is done to prevent a possible situation whereby some templates are amplified preferentially and, after a number of cycles, would thus greatly outnumber those which are more difficult.

It is advisable to have as low an annealing temperature as possible for the sake of producing the most diverse template attainable. On the other hand, compromise is necessary in order to avoid the unspecific binding of primers. The size of the obtained product needs to be carefully checked on an agarose gel of 1.2–1.5%. Non-specific products can usually be seen in electrophoresis as one or more bands of a smaller size than the target. Increasing the annealing temperature by 1–3°C of PCR is suggested.

It is worth mentioning one polymerase which differs from the others, namely, the phi-29 polymerase. With this enzyme, it is possible to amplify DNA from just few microbial cells and even to obtain micrograms of DNA. It is also an extraordinary enzyme in terms of working conditions; the amplification is performed at 30°C, because there is no need to melt double stranded DNA. Reaction takes around 10 to 16 hours and random hexameric primers are used. When employing this enzyme, it is necessary to be aware of its drawbacks. Because of the high efficiency of the phi-29, a careful cleaning procedure needs to precede the reaction setup. Nonetheless, its negative control never works in case of bacteria amplification and a PCR product in a tube in which no template has been applied has to be accepted. This is caused by the fact that the polymerase is obtained from recombinant *E. coli* cells and traces of the host DNA are present in the enzyme solution. This very efficient polymerase catches them as well in PCR and amplifies them. To overcome the results of this artefact, the negative control product can be sequenced and cut from the samples.

It is not only phi-29 polymerase which might result in artefacts, since standard PCR also introduces important biases; the primers might not bind uniformly to the target of interest and, in consequence, some members of the community would amplify either less or not at all [38]. In another step after extraction, we decrease the true genomic diversity of our sample, because, at the very least, there is no universal pair of primers covering all the variants of a given gene.

2.5 Sequencing

Sequencing of an amplified gene marker or, alternatively, of a crude extracted genomic DNA or reverse-transcribed cDNA, can be carried out using shotgun sequencing.

2.5.1 First generation sequencing and clone libraries

This is the traditional sequencing method known as Sanger sequencing, which requires large amounts of DNA. For this reason, the amplified or fragmented DNA is cloned into bacterial plasmids and the transformed host culture is grown on a media to multiply the template. This is a clone library. The procedure makes the method very laborious and, when the price of a single sequence obtained is calculated, it also proves to be an expensive one [25]. Nevertheless, because a 96-well plate of clones per sample is sequenced, the number of sequences per sample is small and the final cost is lower than that of newer sequencing methods for microbial communities. The wide spread presence of Sanger sequencers in laboratories around the world makes the method easily accessible for those who are satisfied with a dominant member community structure. An example of the data analysis workflow for this is shown in Figure 1.

2.5.2 Second generation sequencing

Sanger technology was succeeded by 454 pyrosequencing, which was commercialised by 454 Genomics and, later, by Roche, SOLiD from Life Technologies and Ilumina, formerly known as Solexa. These methods require DNA molecules of a certain length spectra. In the case of shotgun sequencing, the DNA needs to be randomly fragmented. The fragments are ligated with adapters, which helps in the DNA's adhesion to the solid surface on which amplification of unique single molecules takes place, this being beads, for 454-pyrosequncing and SOLiD, or glass flow chambers for Ilumina [39]. Ilumina offers low cost per run and Mb with highest reads number in one run [40]. However because its mid length reads, it is limited to sequencing, where product can be assigned to a known reference. It is particularly useful in counting eg. of functional genes copy numbers. Because SOLiD offers similar sequence mid length reads resequencing and frequency analyses are preferably applicable to it too. Additionally to that SOLiD is known for the lowest error

1. Open the *.abi file in FinchTV program
2. Check the sequence length. Short ones exclude from analysis.
3. Check the “N” nucleotides. If they appear at lower than 800 bases check visually if the base easly recognizable from a peak and change manually
4. Reverse complentary r-primed fragment
5. Save file as *.fasta
6. Make contigs from f and reversed r fragments: CAP3 on line program or optionally ClustalX, BioEdit, BioLign, GeneDoc, BLAST
7. Gaps and chimera check: Mallard program
8. Search the most similar organism sequence: BLAST on line
9. Within samples comparisons: MOTHUR program

Figure 1. Example of clone library data processing.

rate. Because the long sequences reads provides better assemblage and are obtained from 454 in comparison to the other second generation sequencing methods, this is the method, which used to be in preference in sequencing *de novo* and analysis of environmental samples [40,41].

454 and pyroseqencing

The name comes from the pyrophosphate released during nucleotide incorporation. This starts the secondary enzymatic reactions in which light is launched and each nucleotide base is recognised [42-44].

Raw DNA can be a template for a shotgun sequencing version or PCR product can be used for marker gene profiling. For microbial community studies, the latter, in particular, has gained popularity. Constructing an amplicon library can be done by means of several strategies. One is a PCR with the environmental DNA extract run straight forward with forward A and backward B 454-adapters. Because these adapters are heavy and disintegrate easily and PCR can often be low in yield, the second option is a nested PCR. First, a PCR with standard primers amplyfing a target gene is performed and then the amplicons obtained are a subject of a few extra cycles, during which, the sequencing adapters are incorporated.

In addition, barcoded primers can be used, which gives the advantage of pooling the number of samples into one [45]. Each environmental sample is amplifed with a uniquely coded, forward primer-adapter A, which, after pooling and multiplex sequencing, allows the sequences to be assigned to their original environmental sample. It is also possible to use primers coded for different genes and sequence them in parallel in one sample [41]. These ‘tricks of the trade’ decrease reagent usage, costs and time.

When using such degenerated primers, such as barcoded ones for microbial community profiling, one important consideration is that they are provided in quantities which are high as compared to standard PCR protocols in order to make sure that amplification is not limited by the availability of a certain primer variant. It is also important to keep the number of cycles low and, preferably, at no more than twenty to twenty-five cycles. A target product concentration of 5–10 ng μl^{-1} is suitable, but this also depends on the specific PCR conditions and whether any contaminants are inhibiting the reaction. Each sample should be amplified in triplicate to minimise random PCR drift. The PCR product needs to be cleaned from the remaining substrate and primers. Such cleaning is also necessary between the two PCR reactions when nested PCR is used. Cleaning PCR product can be done with silica spinning columns or alternatively, with magnetic beads (Agencourt AMPure XP). The latter can be used for extracting DNA from soil, as well as for cleaning PCR products. Although the Ampure manual, which can be found on line at https://www.beckmancoulter.com/wsrportal/bibliography?docname=Protocol_000387v001.pdf, advises using 1.8 Ampure beads:extract v:v ratio, a ratio of around 0.7 may result in higher template recovery for the cleaning of amplicon libraries from primers. It is advisable to experiment with different ratios and measure the remaining concentration of DNA with a purity control on a gel. Samples for pyrosequencing typically need to be provided in a concentration of above 2 ng μl^{-1} to a total of about 100–200 ng for the sum of all the barcoded libraries. Libraries are preferably stored in water at -20°C, but are quite stable and will last for several weeks at room temperature as well.

Comparison of clone and amplicon library sequencing

In comparison to traditional clone library-based sequencing of microbial communities, pyrosequencing typically results in higher diversity, thanks to the increased sequencing

depth. The protocol for amplicon library sequencing is also more straightforward, while clone library construction is a laborious process in comparison to newer sequencing techniques. Nonetheless, there are two reasons why clone library sequencing is still in use, namely, costs and data processing. It is said that pyrosequencing is cheaper per sequence obtained than clone library sequencing [25]. This makes the whole genome sequencing cheaper. Second generation sequencing is financially demanding, though, since this calculation is arrived at by dividing a large sum per high number of sequences obtained. The method requires extra financing and is more appropriate for those wishing to obtain a deep insight into a community's structure. The quantity of data obtained in second generation sequencing also demands high-throughput bioinformatic methods for processing the data. Just as clone library sequences can be analysed by the researcher himself because the sequences are 700 bp long and are easier to handle, so the analysis of an amplicon library should be carried out by a specialist because of difficulty of constructing contigs and the amount of data involved. *A Bioinformatician's Guide to Metagenomics* [46] or *A primer on Metagenomics* [47] provides a more detailed introduction to second generation sequencing data processing.

2.5.3 RNA (cDNA) sequencing

The sequencing of environmentally derived, reverse-transcribed RNA (cDNA) is often called 'metatranscriptomics', in contrast to the sequencing of genomic DNA, which is known as 'metagenomics'. Shotgun metatranscriptomics includes cDNA derived from mRNA at a native ratio of about 3%. This ratio is typically increased by negative hybridisation of the rRNA, which allows for the observation of gene expression. This is a valuable tool used to predict the metabolic state and functions of the community. Moran's excellent and clear introduction to metatranscriptomics can be reviewed for more information [48].

Retaining the cDNA and sequencing cDNA from the total RNA pool without negative hybridisation has been dubbed 'the Double RNA approach' by Urich *et al*. [49]. It provides information on metabolic activity, together with the active community structure. Analysis of the rRNA faction has two added advantages: 1) No PCR bias is present and 2) data on ribosomal content, which is generally correlated to metabolic activity, from all three domains of life is retained. For a comparison of these methods, see [50].

2.5.4 Shot-gun sequencing

This is sequencing of crude metagenomic extract with no PCR. Shotgun sequencing is aimed at functional profiling rather than community structure analysis and allows to obtain the whole genome structure and metabolic potential of the community being studied. When applied to a highly unequal community, dominated by one or several 'species', it can allow almost the full genome sequences of its members to be assembled. For pyrosequencing, this requires a large quantity of template directly from the extraction, but the quality of the DNA may not be as critical as for Sanger sequencing (see, for example, [51]). A similar endpoint, although a more laborious procedure, would be sorting DNA fragments from extracts by means of Pulse Field Electrophoresis (PFEG), followed by the construction of a fosmid library, which also permits the isolation of almost the full genomes of members of a microbial community.

3. Conclusions

This mini-review was written with a view to disseminating information regarding the newer microbial community analysis methods among a wide range of environmental microbiologists. We targeted microbial communities because the study of these, rather than of populations or single organisms, gives rise to the need for different strategies, including the choice of methods. Even though there are reviews which address metagenomics, they either focus on the theoretical aspects of these methods or retrieve detailed technical considerations suitable for specialist. Our intention was to bridge the gap and, on the one hand, look at the practical aspects with which a researcher may be faced when choosing the proper method for conducting their study, while, on the other hand, still staying with more basic considerations. Ample descriptions of sample preparation and conclusions from laboratory work have been included, since they may well be useful to people instigating work in a molecular laboratory.

Over the last decade, it is the use of clone library sequencing and amplicon library sequencing which have proved to be the most popular and commonly used methods within genetics tools for the study of microbial community structures, with amplicon library sequencing having recently overtaken the older, clone library sequencing. In comparison to Sanger sequencing, second generation sequencing is characterized by a shorter sample preparation time, a lower template quantity needed, the shorter sequences obtained as opposed to the 1000 bp attained from Sanger [42], the higher quantity of data, the necessity of a bioinformatic pipeline, the higher resolution of the community structure, a smaller per-sequence cost yielded and a

higher per-sample cost. The upcoming third generation of sequencing methods will overcome some of the disadvantages of the second generation; the costs promise to be lower [39,40], the sequence length will increase to 1500 bp (for example, Pacific Biosciences), amplification can be dispensed with, thanks to actual single molecule sequencing [41] and quantifiability will be improved. The development of newer sequencing techniques is a fast-moving field and the new, cutting-edge solutions which are constantly being reported are an assurance that the method can only spread into all the biology disciplines.

When used appropriately, sequencing represents a very useful technology for the study of microbial communities. However, one needs to be aware of the limitations and drawbacks; namely, the differing efficiency of genetic material extraction and the biasing and loss of part of the community diversity and composition during extraction, PCR and sequencing, along with a number of other challenges. High-throughput sequencing also produces a great deal of data, which can lead to an analysis consisting of time-consuming steps. The analysis of large amounts of data should preferably be done with the help of specialised bioinformatics. Knowing the challenges and limitations, one can treat the samples in such a way as to avoid as many of them as possible and to provide good quality data.

The different methods are investigations of consecutive steps: 1) DNA-based methods give a static view of present material and ecosystem potential, because of the DNA of dormant microbes such as spores, for example, 2) in contrast to 1), rRNA-based methods investigate the truly living community, metabolising, growing, and so forth; ecosystem selects from the potential community, the community adjusted for the present moment in terms of season, substrate presence, and so on. This is a periodic view of an active community and ecosystem selection, 3) classical ecology methods and experiments, such as, for instance, measuring enzyme activities and processes ratios such as respiration rate, cellulose decomposition showing actual and simultaneous output: functional performance.

The application of molecular methods in environmental biology has opened the door to information that had previously been unavailable and widened our knowledge on microbial communities. One needs to remember, though, that molecular methods do not always provide an unbiased view of the whole picture. Molecular investigation should ideally be supported by classical ecological methods and careful study design, including bioinformatic data analysis.

Acknowledgements

I would like to express my special thanks for the financial support provided, through Individual Mobility Grant No. FSS/2010/II/D3/W0135, by the Scholarship and Training Fund (STF), an initiative financed as under the EEA Financial Mechanism and the Norwegian Financial Mechanism: Iceland, Lichenstein and Norway. My sincere gratitude is owed to Anders Lanzen, Antonio Garcia-Moyano, Ewa Śliwińska and Lise Øvreås for their invaluable support as regards the content of this article.

References

[1] Janssen P.H., Yates P.S., Grinton B.E., Taylor P.M., Sait M., Improved Culturability of Soil Bacteria and Isolation in Pure Culture of Novel Members of the Divisions Acidobacteria, Actinobacteria, Proteobacteria, and Verrucomicrobia, Appl. Environ. Microbiol., 2002, 68, 5, 2391-2396

[2] Nichols D., Cahoon N., Trakhtenberg E.M., Pham L., Mehta A, Belanger A., et al., Use of ichip for high-throughput in situ cultivation of “uncultivable” microbial species, Appl. Environ. Microbiol., 2010, 76, 2445-2450

[3] Stefanowicz A., The Biolog Plates Technique as a Tool in Ecological Studies of Microbial Communities, Polish. Jour. Environ. Stud., 2006, 15, 669-676

[4] Pinkart H.C., Ringelberg D.B., Piceno Y.M., MacNaughton S.J., White D.C., Biochemical Approaches to Biomass Measurements and Community Structure Analysis, In: Hurst C.J., Crawford R.L., Knudsen G.R., McInerney M.J., Stetzenbach L.D. (Eds.), Manual of Environmental Microbiology, ASM Press, Washington D.C., 2002

[5] Olsson P., Signature fatty acids provide tools for determination of the distribution and interactions of mycorrhizal fungi in soil, FEMS Microbiol. Ecol., 1999, 29, 303-310

[6] Frostegård Å., Tunlid A., Bååth E., Use and misuse of PLFA measurements in soils, Soil Biol. Biochem., 2011, 43, 1621-1625

[7] Tyson G.W., Chapman J, Hugenholtz P, Allen E.E., Ram R.J., Richardson P.M., et al., Community structure and metabolism through reconstruction of microbial genomes from the environment, Nature, 2004, 428, 37-43
[8] Simon C., Daniel R., Metagenomic analyses: past and future trends, Appl. Environ. Microbiol., 2011, 77, 1153-1161
[9] Cantarel B.L., Erickson A.R., VerBerkmoes N.C., Erickson B.K., Carey P.A, Pan C., et al., Strategies for metagenomic-guided whole-community proteomics of complex microbial environments, PLoS One, 2011, 6, e27173
[10] Huang W.E., Griffiths R.I., Thompson I.P., Bailey M.J., Whiteley A.S., Raman Microscopic Analysis of Single Microbial Cells, Anal. Chem., 2004, 76, 4452-4458
[11] Harz M., Rösch P., Popp J., Vibrational spectroscopy-a powerful tool for the rapid identification of microbial cells at the single-cell level, Cytometry, 2009, 75A, 104-113
[12] Adamczyk J., Hesselsoe M., Iversen N., Horn M., Lehner A., Nielsen P.H., et al., The Isotope Array, a New Tool That Employs Substrate-Mediated Labeling of rRNA for Determination of Microbial Community Structure and Function, Appl. Environ. Microbiol., 2003, 69, 6875-6887
[13] Okabe S., Kindaichi T., Ito T., MAR-FISH - An Ecophysiological Approach to Link Phylogenetic Affiliation and In Situ Metabolic Activity of Microorganisms at a Single-Cell Resolution, Microbes Environ., 2004, 19, 83-98
[14] Musat N., Halm H., Winterholler B., Hoppe P., Peduzzi S., Hillion F., et al., A single-cell view on the ecophysiology of anaerobic phototrophic bacteria, Proc Natl Acad Sci USA, 2008, 105, 17861-17866
[15] Schippers A., Neretin L.N., Quantification of microbial communities in near-surface and deeply buried marine sediments on the Peru continental margin using real-time PCR, Environ. Microbiol., 2006, 8, 1251-1260
[16] DeSantis T.Z., Brodie E.L., Moberg J.P., Zubieta I.X., Piceno Y.M., Andersen G.L., High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment, Microbial Ecol., 2007, 53, 371-383
[17] He Z., Gentry T.J., Schadt C.W., Wu L., Liebich J., Chong S.C., et al., GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes, The ISME Journal, 2007, 1, 67-77
[18] Wilson K.H., Wilson W.J., Jennifer L., Desantis T.Z., Viswanathan V.S., Kuczmarski T.A., et al., High-Density Microarray of Small-Subunit Ribosomal DNA Probes High-Density Microarray of Small-Subunit Ribosomal DNA Probes, Appl.Environ. Microbiol., 2002, 68, 2535-2541
[19] Hazen T.C., Rocha A.M., Techtmann S.M., Advances in monitoring environmental microbes, Cur. Opin. Biotech., 2012, 24, 1-8
[20] Zwart G., van Hannen E.J., Kamst-van Agterveld M.P., van der Gucht K., Lindström E.S., van Wichelen J., et al., Rapid screening for freshwater bacterial groups by using reverse line blot hybridization, Appl. Environ. Microbiol., 2003, 69, 5875-5883
[21] Nocker A., Burr M., Camper A.K., Genotypic Microbial Community Profiling: A Critical Technical Review, Microb. Ecol., 2007, 54, 276-289
[22] Schütte U.M.E., Abdo Z., Bent S.J., Shyu C., Williams C.J., Pierson J.D., Forney L.J., Advances in the use of terminal restriction fragment length polymorphism (T-RFLP) analysis of 16S rRNA genes to characterize microbial communities, Appl. Microbiol. Biotechnol., 2008, 80, 365-380
[23] Prosser J.I., Replicate or lie, Environ. Microbiol., 2010, 12, 1806-1810
[24] Magurran E.E., McGill B.J., Biological diversity: frontiers in measurement and assessment, Oxford University Press, Oxford, 2011
[25] Delmont T.O., Robe P., Cecillon S., Clark I.M., Constancias F., Simonet P., et al., Accessing the Soil Metagenome for Studies of Microbial Diversity, Appl. Env. Microbiol., 2011, 77, 1315-1324
[26] Ovreas L., Curtis T.P., Microbial diversity and ecology, In: Magurran E.E., McGill B.J. (Eds.), Biological diversity: frontiers in measurement and assessment, Oxford University Press, Oxford, 2011
[27] Simister R.L., Schmitt S., Taylor M.W., Evaluating methods for the preservation and extraction of DNA and RNA for analysis of microbial communities in marine sponges, J. Exp. Mar. Biol. Ecol., 2011, 397, 38-43
[28] Masek T., Vopalensky V., Suchomelova P., Pospisek M., Denaturing RNA electrophoresis in TAE agarose gels, Anal. Biochem., 2005, 336, 46-50
[29] Blagodatskaya E.V., Blagodatskiĭ S.A., Anderson T.H., Quantitative isolation of microbial DNA from the different types of soils of natural and agricultural ecosystems, Microbiology, 2003, 72, 840-846
[30] Aoshima H., Kimura A., Shibutan A., Okada C., Matsumiya Y., Kubo M.,. Evaluation of soil bacterial biomass using environmental DNA extracted by slow-stirring method, Appl. Microbiol. Biotech., 2006, 71, 875-880

[31] Wang Y., Qian P.-Y., Conservative Fragments in Bacterial 16S rRNA Genes and Primer Design for 16S Ribosomal DNA Amplicons in Metagenomic Studies, PLoS One, 2009, 4, e7401
[32] Klindworth A., Pruesse E., Schweer T., Peplies J., Quast C., Horn M., et al., Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies, Nucl. Acid Res., 2012, 1-11
[33] Mao D-P., Zhou Q., Chen Ch-Y., Quan Z-X., Coverage evaluation of universal bacterial primers using the metagenomic datasets, BMC Microbiology, 2012, 12, 66
[34] Walsh P.S., Erlich H.A., Higuchi R., Preferential PCR Amplification of Alleles: Mechanisms and Solutions, Genome Res., 1992, 1, 241-250
[35] Abd-Elsalam K.A., Bioinformatic tools and guideline for PCR primer design, Afr. Jour. Biotech., 2003, 2, 91-95
[36] Chen Z., Zhang Y., Dimethyl sulfoxide targets phage RNA polymerases to promote transcription, Biochem. Biophys. Res. Comm., 2005, 333, 664-670
[37] Rapley R., Enhancing PCR Amplification and Sequencing Using DNA-Binding Proteins, Mol. Biotech., 1994, 2, 295-298
[38] Sipos R., Székely A.J., Palatinszky M., Révész S., Márialigeti K., Nikolausz M., Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis, FEMS Microbiol. Ecol., 2007, 60, 341-350
[39] Mardis E.R., A decade's perspective on DNA sequencing technology, Nature, 2011, 470, 198-203
[40] Glenn T.C., Field guide to next-generation DNA sequencers, Mol. Ecol. Res., 2011, 11, 759-769
[41] Shokralla S., Spall J.L., Gibson J.F., Hajibabaei M., Next-generation sequencing technologies for environmental DNA research, Molecular Ecology, 2012, 21, 1794-1805
[42] Ahmadian A., Ehn M., Hober S., Pyrosequencing: history, biochemistry and future, Clin. chim. acta, 2006, 363, 83-94
[43] Margulies M., Egholm M., Altman W., Attiya S., Bader J.S., Bemben L.A., et al., Genome sequencing in microfabricated high-density picolitre reactors, Nature, 2005, 437, 376-380
[44] Ronaghi M., Karamohamed S., Pettersson B., Uhlén M., Nyrén P., Real-time DNA sequencing using detection of pyrophosphate release, Anal. Biochem., 1996, 242, 84-89
[45] Hamady M., Walker J.J., Harris J.K., Gold N.J., Knight R., Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex, Nat Methods, 2008, 5, 235-237
[46] Kunin V., Copeland A., Lapidus A., Mavromatis K., Hugenholtz P., A bioinformatician's guide to metagenomics, Microbiol. Mol. Biol. Rev., 2008, 72, 557-578
[47] Wooley J.C., Godzik A., Friedberg I., A Primer on Metagenomics, PLoS Comput Biol, 2010, 6, e1000667
[48] Moran M.A., Metatranscriptomics: Eavesdropping on Complex Microbial Communities, Microbe, 2009, 4, 329-335
[49] Urich T., Lanzén A., Qi J., Huson, D.H., Schleper C., Schuster S.C., Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome, PLoS ONE, 2008, 3, e2527
[50] Lanzén A., Jørgensen S.L., Bengtsson M.M., Jonassen I., Ovreas L., Urich T., Exploring the composition and diversity of microbial communities at the Jan Mayen hydrothermal vent field using RNA and DNA, FEMS Microbiol. Ecol., 2011, 77, 577-589
[51] Gilbert M.T., Tomsho L.P., Rendulic S., Packard M., Drautz D.I., Sher A., et al., Whole-genome shotgun sequencing of mitochondria from ancient hair shafts, Science, 2007, 317, 1927-1930